Plan Overview

A Data Management Plan created using DMPonline

Title: (In)Validating User assumptions in Data Integration

Creator:ZF Mouw

Principal Investigator: Andra Ionescu , Zeger Mouw

Data Manager: Andra Ionescu , Zeger Mouw

Affiliation: Delft University of Technology

Template: TU Delft Data Management Plan template (2021)

ORCID iD: 0000-0001-5113-8497

Project abstract:

Often computer science researchers make several assumptions about the workflow, the needs, and the struggles of data professionals, such as data engineers, data scientists, and data analysts. Our research aims to test these assumptions in real life, observe and collect insights via a use case scenario designed to cover the data pipeline of a practitioner.

ID: 134633

Start date: 18-09-2023

End date: 18-03-2024

Last modified: 19-02-2024

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

(In)Validating User assumptions in Data Integration

0. Administrative questions

1. Name of data management support staff consulted during the preparation of this plan.

My faculty data steward, Richard Grimes, has reviewed this DMP on 06 October 2023.

2. Date of consultation with support staff.

2023-10-06

I. Data description and collection or re-use of existing data

3. Provide a general description of the type of data you will be working with, including any re-used data:

Type of data	File format(s)	How will data be collected (for re-used data: source and terms of use)?	Purpose of processing	Storage location	Who will have access to the data
Audio and screen recording.	.mp4	Online interviews conducted using https://jitsi.ewi.tudelft.nl. The screen is recording using local recording applications: QuickTime.	To gain insights into the data integration pipeline in an industry setting.	Project Storage Drive	MSc student, PhD student and the supervisory professors involved in this study.
Transcripts of recordings.	.pdf	The transcripts will be generated using a local instance of the Whisper package from openai (https://github.com/openai/whisper) recommended by the data stewards.	To be able to analyse the content produced by the interviewee.	Project Storage Drive	MSc student, PhD student and the supervisory professors involved in this study.
Name and signature.	.pdf	Informed consent form.	We do not process this data.	TU Delft SURFdrive	MSc student, PhD student and the supervisory professors involved in this study.
Interview questions.	.pdf	We create the interview questions based on the assumtions gathered from the SOTA literature.	This data is our support for the study.	TU Delft SURFdrive	MSc student, PhD student and the supervisory professors involved in this study.

4. How much data storage will you require during the project lifetime?

- 250 GB - 5 TB

We expect to interview at most 20 participants. The size of each recording depends on the quality format of the recording.

II. Documentation and data quality

5. What documentation will accompany data?

- README file or other documentation explaining how data is organised
- Methodology of data collection

The data will be accompanied by a README file which contains information about the steps taken to collect the data.

III. Storage and backup during research process

6. Where will the data (and code, if applicable) be stored and backed-up during the project lifetime?

- Project Storage at TU Delft

The data will be securely store in the Project Storage provided by TUD. The recordings will be deleted after transcribing.

IV. Legal and ethical requirements, codes of conduct

7. Does your research involve human subjects or 3rd party datasets collected from human participants?

- Yes

8A. Will you work with personal data? (information about an identified or identifiable natural person)

If you are not sure which option to select, first ask your Faculty Data Steward for advice. You can also check with the privacy website . If you would like to contact the privacy team: privacy-tud@tudelft.nl, please bring your DMP.

- Yes

We collect minimal personal data such as name and signature, which are used to sign the Informed Consent form. This data will not be used in combination with the recordings, as we do not wish to disclose the identity of the participants.

8B. Will you work with any other types of confidential or classified data or code as listed below? (tick all that apply)

If you are not sure which option to select, ask your Faculty Data Steward for advice.

- No, I will not work with any confidential or classified data/code

9. How will ownership of the data and intellectual property rights to the data be managed?

For projects involving commercially-sensitive research or research involving third parties, seek advice of your Faculty Contract Manager when answering this question. If this is not the case, you can use the example below.

The recordings will only be listened to by the researchers working on this project. The recoding will be transcribed, after which they will be deleted.

The transcriptions will be annonymised.

The datasets underlying the published papers will not be publicly released. During the active phase of research, the project leader from TU Delft will oversee the access rights to data and other outputs. The datasets will not be released publicly at the time of publication of corresponding research papers.

The participation in this study is entirely voluntary, the participants have the right to withdraw at any moment during the interview, in which case we remove their data. The partipicants can withdraw their data until 31st December. After this period the data has been already processed and incorporated in scientific articles. The published materials (thesis, article) will contain statistics about the collected data and insights derived from the data.

10. Which personal data will you process? Tick all that apply

- Signed consent forms
- Data collected in Informed Consent form (names and email addresses)
- Photographs, video materials, performance appraisals or student results

11. Please list the categories of data subjects

Data professionals: data analysts, data engineers, data scientists outside of TUD organisation.

12. Will you be sharing personal data with individuals/organisations outside of the EEA (European Economic Area)?

- No

15. What is the legal ground for personal data processing?

- Informed consent

16. Please describe the informed consent procedure you will follow:

INFORMED CONSENT FORM FOR PARTICIPANTS

Dear participant,

Thank you for participating in this interview! Your input is very much appreciated and very valuable for this research. You are being invited to participate in our research study which aims to understand the data integration pipeline and workflow done by a data professional.

The research
Often computer science researchers make several assumptions about the workflow, the needs, and the struggles of data professionals (e.g., data engineers, data scientists, and data analysts). Our research aims to test these assumptions in real life, observe and collect insights via a use case scenario designed to cover the data pipeline of a practitioner.

Interview structure
The interview takes roughly 60 minutes and consists of several questions and a use case. During the use case, you will be given a small dataset consisting of several tables. Your goal is to augment a given table with more data (e.g., columns, rows) from the dataset, such as the performance of a given ML model increases. You will be asked to complete the task on your own PC using whichever tools you would normally use. The interview will take place online via https://jitsi.ewi.tudelft.nl, and the screen will be recorded using the local application QuickTime.

Data processing
During the interview an audio recording will be made, so that we can listen to your answers and process the data. Additionally, a screen recording will be made, as we are interested in is observing your behaviour and decision-making process. These recordings will only be listened to and viewed by the researchers working on this project, who are listed below. The recordings will be deleted after the data is transcribed. Your consent form and recording will never be processed in combination. All the data will be stored securely in the Research Project Drive provided by TU Delft. The only personal data collected will be your name and signature on this consent form. Any other personal data will not be collected during this study.

As with any data collection activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimise any risk by securely storing your answers' digital copies in the Research Project Drive provided by TU Delft. Although the video will be deleted as soon as transcripts have been made, it remains the participant's responsibility to ensure that no personal or confidential data is exposed in the screen recording.

Permission
Your participation in this study is entirely voluntary. You have the right to withdraw at any time without giving a reason, in which case your data will be deleted. You can withdraw until 31st December 2023. From this date onwards, your data will be already

processed and included in the research materials. You are free to ask any questions any time during the session.

More information
If you want to know more about this research, or have any questions, you can always contact us.

Research team
Responsible researcher: Zeger Mouw
Email: z.f.mouw@student.tudelft.nl

Supervisory researcher: Andra Ionescu
Email: a.ionescu-3@tudelft.nl

Supervisory professor: Fenia Aivalogou
Email: e.aivaloglou@tudelft.nl

Supervisory professor: Asterios Katsifodimos
Email: a.katsifodimos@tudelft.nl

Please tick the boxes which you understand and give your consent to:

PLEASE TICK THE APPROPRIATE BOXES	**Yes**	**No**
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information dated [__/__/____], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time (until 31stDecember 2023) without having to give a reason.	☐	☐
3. I understand that taking part in the study involves:	☐	☐
- Answering questions and working with open-data for a data integration use case. - The recording of voice. The recording will be transcribed into text and your personal identifiable information will be removed (e.g., name, signature). - The recordings will be securely stored in the Research Project repository.		
4. I understand that the study will end after a maximum allocated time of 1 (one) hour, or when requested at any time.		
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
5. I understand that taking part in the study involves the following risks which will be mitigated by:	☐	☐
- You may experience mental discomfort such as frustration or anxiety during the session, and you can ask to stop participating at any point.		
6. I understand that taking part in the study also involves collecting specific personally identifiable information (PII), such as you name and signature in this consent form, and associated personally identifiable research data (PIRD), such as the voice recordings.	☐	☐
7. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach:	☐	☐
- The audio recording will be securely stored in the Research Project Drive repository. - The recording will be transcribed in text while the personally identifiable information will be removed.		

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
8. I understand that personal information collected about me that can identify me, such as name and signature, will not be shared beyond the study team.	☐	☐
9. I understand that the (identifiable) personal data I provide will be destroyed when the project ends (January 2024).	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
10. I understand that after the research study the de-identified information I provide will be used for conference/journal article publication, master thesis and PhD thesis.	☐	☐
11. I agree that my responses, views or other input can be quoted anonymously in research outputs.	☐	☐
12. I understand that I can request my data be withdrawn from the study at any time during the study until 31st December 2023. After this time, data I have provided will have been processed by the research team and disseminated such that it will no longer be possible to withdraw.		
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
13. I give permission for the de-identified data (such as anonymised quotes) that I provide to be archived in Research Project Drive repository so it can be used for future research and learning.	☐	☐
14. I understand that access to this repository is private and only accessible for the responsible researcher.	☐	☐

Signatures
_____ _____ _____ Name of participant [printed] Signature Date
I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting. _____ _____ _____ Researcher name [printed] Signature Date

17. Where will you store the signed consent forms?

- Same storage solutions as explained in question 6

18. Does the processing of the personal data result in a high risk to the data subjects?

If the processing of the personal data results in a high risk to the data subjects, it is required to perform a Data Protection Impact Assessment (DPIA). In order to determine if there is a high risk for the data subjects, please check if any of the options below that are applicable to the processing of the personal data during your research (check all that apply).
If two or more of the options listed below apply, you will have to complete the DPIA. Please get in touch with the privacy team: privacy-tud@tudelft.nl to receive support with DPIA.
If only one of the options listed below applies, your project might need a DPIA. Please get in touch with the privacy team: privacy-tud@tudelft.nl to get advice as to whether DPIA is necessary.
If you have any additional comments, please add them in the box below.

- None of the above applies

22. What will happen with personal research data after the end of the research project?

- Personal research data will be destroyed after the end of the research project

V. Data sharing and long-term preservation

27. Apart from personal data mentioned in question 22, will any other data be publicly shared?

- All other non-personal data (and code) produced in the project
- All other non-personal data (and code) underlying published articles / reports / theses

We collect audio and screen recording of the participants with the scope of analyzing it and derive insights. The recordings are transcribed, at which point, the recordings can be removed. The transcripts are used by our team to conduct the current research project.
Only the anonymous aggregated results that are produces will be publicly shared at the end of the project.
Besides the aggregated results we can share the designed use case, the accompanying interview questions.

29. How will you share research data (and code), including the one mentioned in question 22?

- All anonymised or aggregated data, and/or all other non-personal data will be uploaded to 4TU.ResearchData with public access

31. When will the data (or code) be shared?

- As soon as corresponding results (papers, theses, reports) are published

VI. Data management responsibilities and resources

33. Is TU Delft the lead institution for this project?

- Yes, the only institution involved

34. If you leave TU Delft (or are unavailable), who is going to be responsible for the data resulting from this project?

Asterios Katsifodimos

a.katsifodimos@tudelft.nl
Assistant Professor in Web Information Systems research group.

Fenia Aivaloglou
e.aivaloglou@tudelft.nl
Assistant Professor in Web Information Systems research group.

35. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

4TU.ResearchData is able to archive 1TB of data per researcher per year free of charge for all TU Delft researchers. We do not expect to exceed this and therefore there are no additional costs of long term preservation.